Williams, T., Szollosi, G. J., Spang, A., Foster, P. G., Heaps, S. E., Boussau, B., Ettema, T. J. G., & Embley, T. M. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(23), E4602–E4611. https://doi.org/10.1073/pnas.1618463114

Integrative modeling of gene and genome evolution roots the archaeal tree of life

Tom A. Williams[a,b,1], Gergely J. Szöllősi[c,2], Anja Spang[d,2], Peter G. Foster[e], Sarah E. Heaps[b,f], Bastien Boussau[g], Thijs J. G. Ettema[d], and T. Martin Embley[b]

[a]School of Earth Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom; [b]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom; [c]MTA-ELTE Lendület Evolutionary Genomics Research Group, 1117 Budapest, Hungary; [d]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden; [e]Department of Life Sciences, Natural History Museum, London SW7 5BD, United Kingdom; [f]School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom; and [g]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, F-69622 Villeurbanne, France

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved April 24, 2017 (received for review November 7, 2016)

A root for the archaeal tree is essential for reconstructing the metabolism and ecology of early cells and for testing hypotheses that propose that the eukaryotic nuclear lineage originated from within the Archaea; however, published studies based on outgroup rooting disagree regarding the position of the archaeal root. Here we constructed a consensus unrooted archaeal topology using protein concatenation and a multigene supertree method based on 3,242 single gene trees, and then rooted this tree using a recently developed model of genome evolution. This model uses evidence from gene duplications, horizontal transfers, and gene losses contained in 31,236 archaeal gene families to identify the most likely root for the tree. Our analyses support the monophyly of DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea), a recently discovered cosmopolitan and genetically diverse lineage, and, in contrast to previous work, place the tree root between DPANN and all other Archaea. The sister group to DPANN comprises the Euryarchaeota and the TACK Archaea, including *Lokiarchaeum*, which our analyses suggest are monophyletic sister lineages. Metabolic reconstructions on the rooted tree suggest that early Archaea were anaerobes that may have had the ability to reduce CO_2 to acetate via the Wood–Ljungdahl pathway. In contrast to proposals suggesting that genome reduction has been the predominant mode of archaeal evolution, our analyses infer a relatively small-genomed archaeal ancestor that subsequently increased in complexity via gene duplication and horizontal gene transfer.

evolution | phylogenetics | Archaea

The Archaea are one of the primary domains of cellular life (1). In addition to the classically defined Euryarchaeota and Crenarchaeota (1), the scope of archaeal diversity has been dramatically expanded in recent years by the discovery of major new lineages using traditional and molecular methods. These lineages are of major ecological and evolutionary significance and include the Thaumarchaeota (2, 3), ammonia oxidizers found in soils and the open ocean, where they play a critical role in the global nitrogen cycle (3); the DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea) Archaea, a diverse group with small cells and genomes, whose reduced metabolic repertoires suggest that they may be symbionts or parasites of other prokaryotes (4, 5); and the "Asgard" Archaea, the closest archaeal relatives of eukaryotes described to date (6, 7), whose phylogenetic position and gene content are key to ongoing debates about eukaryote origins. In recent years, phylogenetic analyses have supported a clade uniting the Thaumarchaeota, Crenarchaeota, Aigarchaeota, and Korarchaeota that has been informally named the "TACK" Archaea (8) or "Proteoarchaeota" (9). The deeper relationships between the major archaeal lineages, and the root of the archaeal tree, remain matters of debate, however. Resolving these questions is important for understanding the origins and evolution of the Archaea, and also for testing hypotheses about the prokaryote-to-eukaryote transition, one of the major unsolved problems in biology.

Recently published analyses (9, 10) used a bacterial outgroup to root the archaeal tree, based on the assumption that the root of the universal tree lies between the two prokaryotic domains or within the Bacteria (11–15). Although outgroup rooting is a widely used approach, it has at least two major difficulties in this context. First, the analysis is restricted to a set of ∼30–70 genes conserved between Bacteria and Archaea that comprises only a small fraction (2–3%) of a typical archaeal genome. Second, tree-based analyses at this depth are plagued by a phylogenetic artifact known as long branch attraction (16, 17); the evolutionary process along the long branch joining the two domains is difficult to model and can induce errors in the resolution of the deepest branches within the in-group. Despite their broadly similar datasets and analytical approaches, previous analyses have reached different conclusions regarding the position of the archaeal root (9, 10, 18). Petitjean et al. (9) rooted the tree between the Euryarchaeota and the TACK Archaea, whereas Raymann et al. (10) placed the root between most of the Euryarchaeota and a clade comprising the TACK Archaea, the Thermococcales, and the cluster I methanogens (a euryarchaeotal clade comprising *Methanococcus*, *Methanothermobacter*, and their relatives). A similar result—euryarchaeote and methanogen paraphyly—was reported by Foster et al. (18). These results provide contrasting predictions about the nature of the ancestral archaeon, given that a root bipartition with methanogens on both sides suggests that the earliest Archaea were

Significance

The Archaea represent a primary domain of cellular life, play major roles in modern-day biogeochemical cycles, and are central to debates about the origin of eukaryotic cells. However, understanding their origins and evolutionary history is challenging because of the immense time spans involved. Here we apply a new approach that harnesses the information in patterns of gene family evolution to find the root of the archaeal tree and to resolve the metabolism of the earliest archaeal cells. Our approach robustly distinguishes between published rooting hypotheses, suggests that the first Archaea were anaerobes that may have fixed carbon via the Wood–Ljungdahl pathway, and quantifies the cumulative impact of horizontal transfer on archaeal genome evolution.

Author contributions: T.A.W., T.J.G.E., and T.M.E. designed research; T.A.W., G.J.S., A.S., P.G.F., S.E.H., and B.B. performed research; G.J.S. and B.B. contributed new reagents/analytic tools; T.A.W., G.J.S., A.S., P.G.F., S.E.H., and B.B. analyzed data; and T.A.W., A.S., T.J.G.E., and T.M.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: tom.a.williams@bristol.ac.uk.

[2]G.J.S. and A.S. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1618463114/-/DCSupplemental.

methanogens. The ancestral state is more ambiguous under the Petitjean et al. (9) root, because methanogenesis could have evolved along the branch leading to the common ancestor of the Euryarchaeota. Inferences about ancestral archaeal physiology have added significance under hypotheses in which the Archaea and Bacteria represent the two primary domains of cellular life (19), because they would also inform ideas about how the first cells evolved and diversified on the early Earth (20, 21).

Here we used a method of rooting the archaeal tree that does not depend on an outgroup and that uses much more of the available genomic data for the root inference. Compared with previous work, our analyses also incorporate an expanded sampling of archaeal diversity, including the DPANN Archaea (4, 5)—a cosmopolitan, genetically diverse, and ecologically important lineage of uncultivated Archaea—and *Lokiarchaeum* (6), a representative of the Asgard Archaea (7). We first combined protein concatenation and a supertree approach using 3,242 single gene trees to resolve a consistent unrooted archaeal topology, and then inferred a root for this tree using a probabilistic gene tree-species tree reconciliation model that integrates information from the evolutionary history of 31,236 archaeal gene families. In addition to providing a root inference, this model-based approach also allowed us to infer properties of the last archaeal common ancestor (LACA), including its genome size and potential metabolism. The reconstructions provide new information about the tempo and mode of genome evolution affecting different Archaea, including estimates of the contributions made by horizontal transfer and lineage-specific evolution to major ecological transitions across the archaeal tree.

Results and Discussion

Identifying a Consensus Unrooted Topology for the Archaea with Concatenated Proteins and Multigene Supertrees. We used the OMA algorithm (22) to identify single-copy orthologs on 62 archaeal genomes sampled from across the known diversity of the domain. Our sample included 21 single-cell genomes and metagenomic bins from uncultivated lineages, which are now known to represent some of the most abundant and ecologically important Archaea (4, 5). We filtered candidate marker genes to remove potential horizontal gene transfers (HGTs) (*Materials and Methods*) and inferred a concatenated protein phylogeny (Fig. 1) for a Dayhoff-recoded (23) supermatrix comprising 45 proteins under the CAT+GTR (generalized time-reversible) model, the best-fitting evolutionary model. To complement the supermatrix analysis, we used a multigene supertree approach, matrix representation with parsimony (MRP) (24–27), which integrates the phylogenetic signal for vertical descent from a much broader sample of genes than can be accommodated by concatenation alone (*Materials and Methods*). The supertree inferred by MRP fitted to a dataset of 3,242 single gene trees (*SI Appendix*, Fig. S1) is in good agreement with the concatenation tree topology (Fig. 1), providing a robust phylogenetic backbone for rooting analysis and suggesting strong vertical signals in the data. The main difference between the two trees is the position of the Thermococcales, which emerge at the base of a clade comprising the TACK Archaea and *Lokiarchaeum* in the concatenation tree but at the base of the Euryarchaeota in the supertree. This clade has been difficult to place in previous analyses, and both of the positions that we recovered have been supported by previous work (9, 10, 18, 28). We evaluated both positions for the Thermococcales in our rooting analysis.

The unrooted phylogeny contains three major clans (29), or potential clades defined by a single split on the tree. These correspond to (*i*) a metabolically diverse assemblage comprising the TACK (8) Archaea—Thaumarchaeota (2), Aigarchaeota (30), Crenarchaeota (31), Korarchaeota (32)—and the recently discovered Lokiarchaeota (6), which emerges as the sister group of TACK; (*ii*) the core Euryarchaeota, comprising the methanogenic Euryarchaeota and their relatives (31), with the

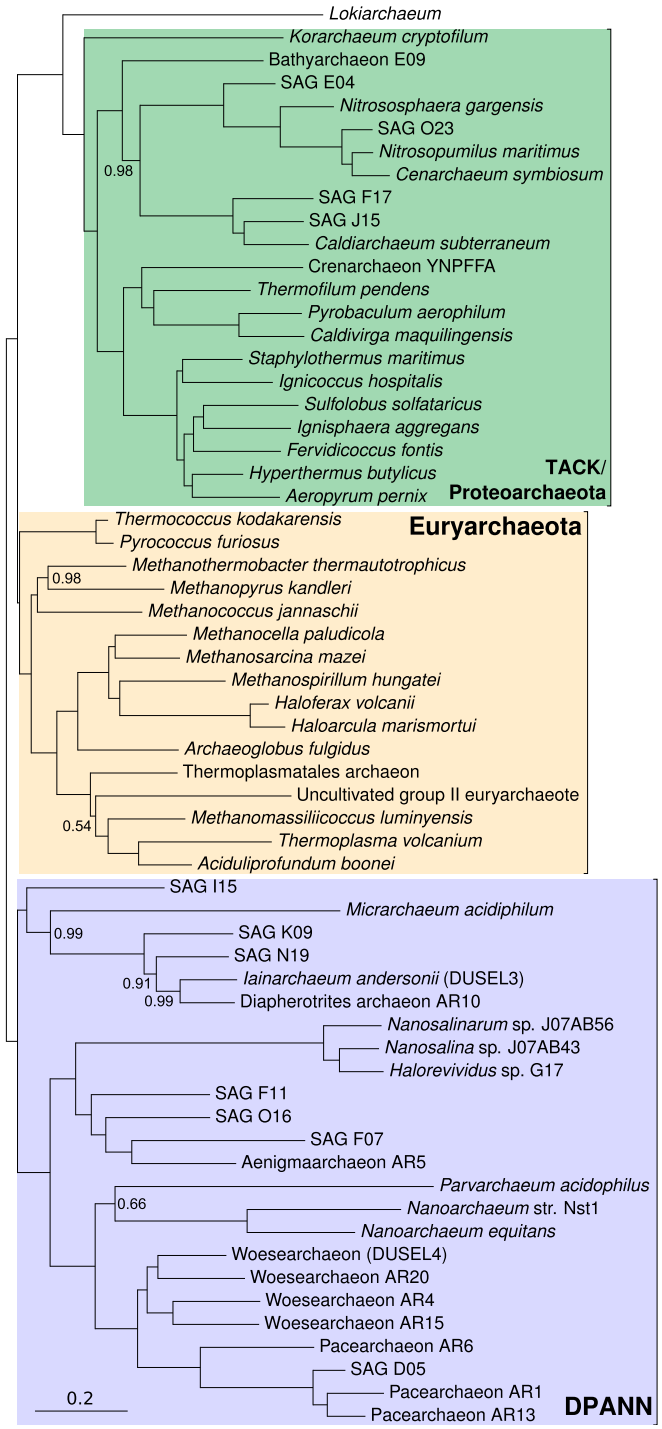

Fig. 1. A rooted tree of the Archaea. This rooted phylogeny summarizes inferences from analyses of a concatenation of 45 protein-coding genes under CAT+GTR, an MRP supertree of 3,242 single-copy, lineage-specific archaeal gene families, and DTL modeling of archaeal gene family evolution using the ALE method. The concatenation and supertree analyses recovered the same unrooted topology for all but the Thermococcales, which grouped at the base of the TACK/*Lokiarchaeum* clade in the concatenated protein analysis but at the base of the Euryarchaeota in the supertree. We obtained significantly better likelihoods for a Thermococcales+Euryarchaeota clade from DTL modeling, and that is the topology depicted here. Support values are Bayesian PPs from the CAT+GTR+Dayhoff analysis, and branch lengths are expected numbers of substitutions per site under CAT+GTR+Dayhoff. The tree is rooted according to the ML root position obtained in the DTL analysis, as discussed in the text.

exception of the Thermococcales; and (*iii*) the DPANN Archaea (4), comprising *Nanoarchaeum* and its relatives. Many of the DPANN Archaea described to date have small cells (<1 μm) (5) and reduced genomes in which many core metabolic pathways are incomplete (5). The first DPANN to be characterized, *Nanoarchaeum equitans*, is obligately dependent on the crenarchaeote *Ignicoccus hospitalis* for growth (33), and other members of the group also have been observed in direct contact with larger archaeal cells (34), suggesting that symbiotic or parasitic lifestyles may be a common feature of the DPANN lineage. Nevertheless, genome analyses suggest that at least some members of the group might be capable of a free-living lifestyle (35).

Are the DPANN Archaea a Clan? Our phylogeny is in agreement with recent published analyses in recovering the clanhood of TACK (8, 28, 36, 37) and the core Euryarchaeota (5, 9, 10), and supports the placement of *Lokiarchaeum* at the base of the TACK phylum (6). DPANN clanhood is in agreement with some recent reports (4, 5, 38), but has been challenged (9) on the grounds that the high rates of sequence evolution shared by some DPANN lineages make them vulnerable to long-branch attraction (LBA) (39, 40). DPANN clanhood raises unsettling parallels with early molecular phylogenies of the eukaryotes, in which fast-evolving parasitic lineages were drawn to the base of the tree by LBA (41). Published DPANN phylogenies have shown conflicting results (2, 5, 9, 42), although analyses using the CAT+GTR model, which may be less susceptible to LBA than simpler methods (43), have recovered DPANN monophyly (38). Owing to this uncertainty, recent analyses of the archaeal root have excluded DPANN (10) or have included only a subset of sequenced lineages (9), on the grounds that their presence would interfere with the overall resolution of the tree. The limitation of this approach is that DPANN lineages are ecologically important (44) and represent a substantial proportion of known archaeal diversity (4, 5). Therefore, any analysis of the archaeal root that does not account for their origins is necessarily incomplete.

We performed a series of tests designed to investigate whether DPANN clanhood could be attributed to LBA. We first recoded the alignment into the four Dayhoff categories (23, 45), which made the data easier to model by reducing both compositional heterogeneity and substitutional saturation (46). Analysis of this recoded matrix under CAT+GTR, one of the best phylogenomic models for ameliorating the effects of LBA (43), recovered DPANN clanhood with maximal posterior support [posterior probability (PP) = 1; Fig. 1]. We then selectively removed the longest-branching DPANN lineages from the analysis (*Materials and Methods* and *SI Appendix*, Fig. S2), and again obtained maximal support for DPANN clanhood, although Pacearchaeota now clustered within Woesearchaeota with moderate support (PP = 0.89). We also reanalyzed the original alignment after applying a more stringent approach to identify and remove fast-evolving sites (the BLOSUM62 matrix in BMGE) (47), which are considered the sites most susceptible to LBA artifacts (16); support for DPANN clanhood was unchanged (*SI Appendix*, Fig. S3).

Next, we reasoned that if the DPANN lineages were being artifactually drawn to the base of the tree because of LBA, then an analysis of DPANN alone might not reveal the same in-group topology as that seen in the full analysis, including the euryarchaeotal and TACK outgroups (17). Artifacts of this type have previously been observed in analyses of within-eukaryote relationships, whereby fast-evolving eukaryotes that were drawn toward the prokaryotic outgroup were recovered in the expected position in a eukaryote-only analysis (fig. S62 in ref. 48). However, a CAT+GTR analysis of the DPANN portion of the concatenation alone resulted in a topology compatible with that of the overall tree (*SI Appendix*, Fig. S4), with the exception of the position of *Haloredivivus*, which exchanges with its nearest neighbor in the DPANN-only reanalysis at PP = 0.88. We also considered the possibility that DPANN clanhood might be an artifact of nonrandom gene representation in the supermatrix, which potentially could lead to systematic error (49). Because many DPANN lineages were represented by an incomplete metagenomic bin, they often contained more gene absences (and thus gaps) in the supermatrix compared with other Archaea (DPANN gene representation ranging from 4 to 40 genes; median, 28 genes). To evaluate the impact of this gene representation bias on our analyses, we subsampled the original dataset, selecting the most complete DPANN genomes (10 genomes, including at least one genome from each major DPANN sublineage) and the most widely conserved genes (25 genes) to produce a supermatrix in which gene representation was equal across Euryarchaeota, TACK, and DPANN. Analysis of this supermatrix under CAT+GTR resulted in a topology (*SI Appendix*, Fig. S5) almost identical to that inferred from the full concatenation (Fig. 1). In particular, support for both the in-group relationships within DPANN and the clanhood of the group as a whole were identical to those seen in the original analysis.

Finally, and because DPANN Archaea tend to have above-average evolutionary rates, we considered the possibility that the apparent clanhood of the group as a whole is the result of LBA between the stems leading to each distinct sublineage. Thus, we investigated the behavior of individual sublineages in a series of concatenated protein analyses from which all other DPANN Archaea had been removed (*SI Appendix*, Figs. S6–S12). The idea is that if DPANN are monophyletic, then their constituent lineages should each individually connect to the same point on a subtree containing only Euryarchaeota and the TACK/*Lokiarchaeum* clade. In these analyses, the Diapherotrites, Aenigmaarchaeota, and Woesearchaeota lineages—composing just over one-half (14 of 24) of sampled DPANN lineages—fell between the Euryarchaeota and TACK/*Lokiarchaeum* clans, as in the full analysis; however, the remaining DPANNs grouped at the base of the Euryarchaeota, either with the Thermococcales (*Nanoarchaeum*) or within the cluster 1 methanogens (Nanohaloarchaeota, Pacearchaeota, and the solitary *Parvarchaeum*). The difficulties in finding stable positions for single DPANN lineages including *Nanoarchaeum*, Nanohaloarchaeota, and *Parvarchaeum* is already clear from comparing trees in previously published work (2, 10, 38, 42, 50). The results of single-lineage analyses are also difficult to compare with the full analysis, because there is no principled statistical framework within which to evaluate whether attempting to place DPANN lineages individually ameliorates or aggravates potential phylogenetic artifacts, such as LBA. Better taxonomic sampling has been shown to improve phylogenetic inference (51–53), and there is no posterior support for these alternative placements from any of our analyses in which DPANN monophyly was tested directly, including our supertree analysis and the series of supermatrix analyses performed with methods commonly used to ameliorate LBA (23, 43, 54) (*SI Appendix*, Figs. S2–S5). Nonetheless, we considered both possibilities—monophyletic and polyphyletic DPANN—and also performed an analysis in which all DPANN lineages were excluded in our subsequent rooting and gene content analyses.

Using a Bacterial Outgroup to Root the Archaea. Recent work using a bacterial outgroup to root the Archaea has recovered a root either between Euryarchaeota and TACK (9) or within the Euryarchaeota (10). Our own outgroup rooting analysis using 29 universally conserved protein-coding genes and the CAT+GTR model (*SI Appendix*, Table S2 and Fig. S13) did not robustly distinguish between these two hypotheses. We obtained weak to moderate posterior support for the exclusion of four clades from the root (DPANN, TACK/*Lokiarchaeum*+Thermococcales, core Euryarchaeota, and cluster I methanogens), whereas the basal split within the Archaea was unresolved. The outgroup approach allows the addition of a priori rooting information to trees inferred

under standard models of sequence evolution, which do not directly infer the root (55, 56). However, outgroup rooting is known to be problematic when the outgroup is distantly related to the ingroup (16, 17, 57), as is the case when one cellular domain is used to root another. The length of the branch leading to the outgroup is particularly striking in our analysis (*SI Appendix*, Fig. S14), where the bacterial stem was predicted to have experienced 4.79 substitutions per site, compared with a mean of 0.192 (range, 0.0157–0.546) for within-domain branches. The use of long outgroup branches is a general problem that has contributed to disagreements about the archaeal root as well as about the roots of other major radiations (58–61), motivating a search for alternative rooting methods.

Bringing More Data to Bear on the Archaeal Root. The use of gene duplications to root major clades has a venerable history in molecular evolution (45), particularly for resolving the root of the tree of life (11–13, 62–64). More recently, it has been appreciated that gene gains, losses, and horizontal transfers also contain information about the root of a species tree that can be integrated using probabilistic gene tree-species tree reconciliation approaches (65–67). We used a recently developed method known as amalgamated likelihood estimation (ALE) (67) to calculate gene family likelihoods for each of the 31,236 homologous gene families encoded by our sample of 62 archaeal genomes, under a set of candidate root positions on the archaeal species tree (Fig. 1) corresponding to published rooting hypotheses as well as a selection of other plausible rooting positions, such as between each of the major lineages (*SI Appendix*, Tables S3–S5). We also evaluated a species tree in which the DPANN were polyphyletic (*SI Appendix*, Fig. S15), as has been suggested by some single-lineage supermatrix analyses. Different roots on the species tree imply different scenarios of gain, duplication, transfer, and loss for the gene families observed on modern genomes (Fig. 2), and because of this they have different likelihoods under the model. The rates of gene duplication, transfer, and loss (DTL) were inferred from the data using maximum likelihood (ML) optimization, and ALE incorporates uncertainty in the underlying gene trees using conditional clade probabilities (68, 69). This means that poorly supported disagreements between the species tree and the gene trees do not unduly affect estimates of the number of DTL events.

Using an approximately unbiased test (70) to establish a confidence set from our analyses for the archaeal root at $P > 0.05$, we were able to reject all but a single root (*SI Appendix*, Table S3), the root between DPANN and a clade comprising the Euryarchaeota and TACK/*Lokiarchaeum* lineages (Fig. 3*A*). We obtained significantly higher likelihoods using rooted trees in which the Thermococcales were placed at the base of the Euryarchaeota rather than with the TACK Archaea, in agreement with our supertree (*SI Appendix*, Fig. S1) and some previous analyses (9, 10, 48). The tree in which DPANN were polyphyletic had the worst likelihood score of the trees considered and was rejected at $P = 4 \times 10^{-4}$. As an additional control against potential LBA artifacts that might result from the inclusion of DPANN, we also repeated the analysis, including inference of the underlying single-gene trees, without DPANN. The confidence set for this reduced analysis consisted of two trees, a tree placing the root between Euryarchaeota and the TACK/*Lokiarchaeum* lineage consistent with the full analysis, and a tree in which the root was placed on the branch leading to *Lokiarchaeum* (*SI Appendix*, Table S4). Although the *Lokiarchaeum* root could not be rejected in the reduced analysis, it was rejected by the full dataset, and it is not supported by our outgroup rooting analysis (*SI Appendix*, Fig. S13) or by published phylogenetic and comparative genomic analyses that group *Lokiarchaeum* with other "Asgard" Archaea within the TACK lineage (6, 7, 71). None of our analyses based on patterns of gene DTL provided any support for a root within a paraphyletic Euryarchaeota (10, 18).

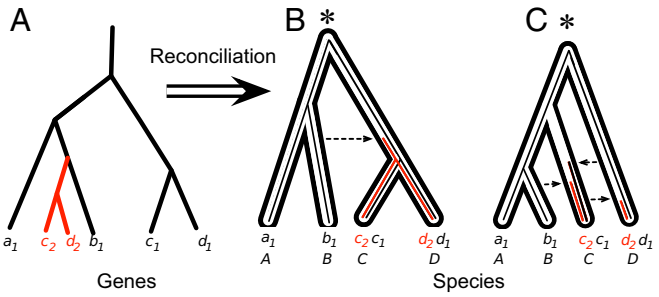

Fig. 2. Using gene DTL to root the species tree. Different roots (denoted by asterisk) on the species tree imply different scenarios of gene family evolution, and thus lead to different gene family likelihoods under the probabilistic gene tree-species tree reconciliation model implemented in ALE (67); here we provide a simple illustration of the approach. (*A*) The evolutionary history of a gene family present in two copies in species C and D, but only a single copy in A and B. Solid lines indicate the branches of the inferred gene tree, and red highlights represent discord with the species tree. The number of gene transfers needed to explain this gene tree depends on the root of the species tree. (*B* and *C*) A root between species AB and CD would require one transfer (*B*), but a root between ABC and D would require three transfers (*C*), providing some support for the root depicted in *B*. Other reconciliations (e.g., gene duplications above the root followed by a series of losses) are also possible; ALE integrates over these possibilities to calculate a likelihood for each gene family under each root position. Rooting hypotheses can then be statistically distinguished from one another based on these likelihoods.

DTL modeling approaches have been developed only recently, and their limitations are still being evaluated. To determine the robustness of our results, we performed a series of sensitivity and simulation analyses, which analyses indicated that our root inference is robust to high rates of horizontal transfer and variation in species sampling among gene families, and that our method robustly recovers the true root on simulated data (*SI Appendix*). An additional source of DTL error might be a kind of "small genome attraction," in which the model favors a root that divides smaller from larger genomes on the tree. To investigate whether this might have been responsible for the support for basal DPANN, we repeated the rooting analysis using the 2,492 gene families that included at least one sequence from a DPANN archaeon. The 5% confidence set for the analysis of this reduced dataset contained only two rooted trees (*SI Appendix*, Table S5) and in both cases the root was placed between DPANN and all other Archaea. The difference between the two trees again lay in the position of the Thermoccocales, which were placed at the base of either the TACK Archaea or the Euryarchaeota. All of our sensitivity analyses agreed with the full data set in rejecting a root on *Lokiarchaeum* or among the Euryarchaeota.

Reconstructing Ancestral Archaeal Metabolisms. Our DTL analysis provides an inference of the history of gene family evolution, including estimates of ancestral genome content (Fig. 3). To reconstruct ancestral metabolisms, we assigned functional annotations to the genes predicted to be present at each internal node on the tree, and mapped these onto core archaeal metabolic pathways (6). It is important to realize that these reconstructions are necessarily incomplete, because it is possible to reconstruct the history only of gene families that have survived to the present day in at least one of the sampled genomes. Moving back in time, the probability that genes on ancestral genomes survived to the present day decreases, and we estimate that 41% of the gene families that were present on the genome of the archaeal common ancestor have since gone extinct (*Materials and Methods*). These extinction probabilities can be used to correct ancestral genome size estimates (see below), although the functions of the extinct genes remain unknown.

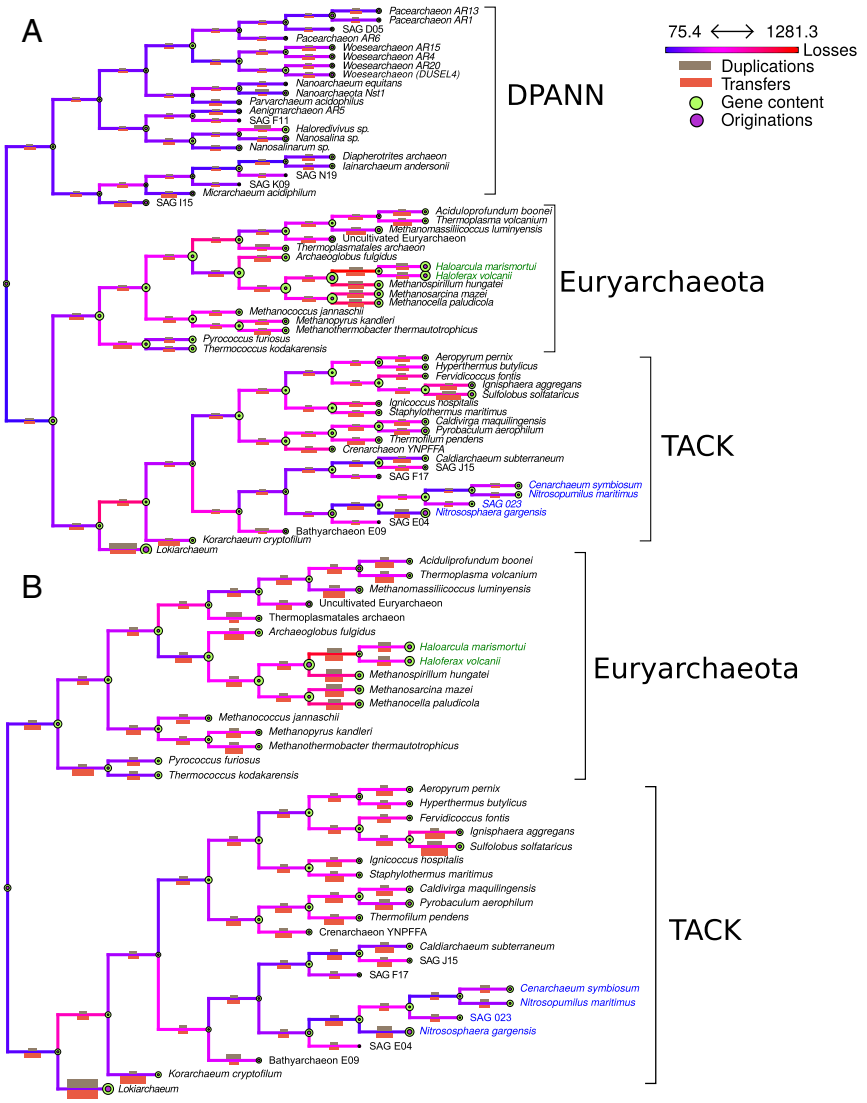

Fig. 3. An ML reconstruction of archaeal gene family evolution. We used the DTL model implemented in ALE (67) to perform gene tree-species tree reconciliation using the rooted tree shown in Fig. 1 and the set of homologous gene families from our sample of archaeal genomes. The diameters of the circles at each node are proportional to inferred gene content (1,090 gene families at the root) and number of originations, or new genes. Branch colors denote number of gene losses, and the areas of the bars above and below each branch correspond to numbers of gene duplications and HGTs. The complete data underlying this figure are provided in *SI Appendix*, Table S8. The analysis was performed with (*A*) and without (*B*) the inclusion of the DPANN lineages. In contrast to scenarios in which a complex archaeal common ancestor gave rise to modern lineages by streamlining (89), fitted DTL models imply a common ancestor whose genome was moderately smaller than modern lineages, with an ongoing process of genome expansion via gene duplications, de novo gene origination, and HGTs throughout archaeal evolution. The Haloarchaeota (green) and the Thaumarchaeota (blue) are the two stem lineages that have experienced the greatest number of gene acquisitions, whether by de novo innovation or by HGT.

Our analyses (Fig. 4) suggest that the LACA was an anaerobe that may have been able to fix CO_2 to acetyl-CoA via the Wood–Ljungdahl pathway and to subsequently generate acetate and ATP using an acetyl-CoA synthetase (arCOG01340) (*SI Appendix*). It is challenging to map genes to deep nodes in the phylogeny with high probability, and our reconstruction did not allow us to determine whether the electron donor for this reaction was organic or inorganic. Our mapping of the Wood–Ljungdahl pathway to the deepest nodes of the archaeal tree is in agreement with biochemical arguments and recent analyses using different methods (72, 73) that have suggested that the reduction of CO_2 with H_2 to produce organic compounds was central to the metabolism of an anaerobic last common ancestor of Bacteria and Archaea. Our analysis also suggests that the LACA had most of the modern archaeal transcription, translation, and DNA replication machineries, components of the exosome and proteasome, a secretion system, and some of the key genes for synthesizing archaeal ether lipids (*SI Appendix*, Figs. S16 and S17 and Table S6). The presence of lipids with a glycerol-1-phosphate backbone is often considered a hallmark of the Archaea and has been used to argue against the idea that eukaryotes originated from the endosymbiosis of the mitochondrial ancestor within an Archaeon (19, 20, 74, 75). Interestingly, we could not confidently map the origin of the key enzyme glycerol-1-phosphate dehydrogenase, which synthesizes the glycerol-1-phosphate backbone of canonical archaeal lipids, to the base of the Archaea, because it is absent from the published genomes or metagenomes of group II/III euryarchaeota, several members of the DPANN, and *Lokiarchaeum* (76). At least in the case of the Euryarchaeota and *Lokiarchaeum*, these absences likely reflect recent lineage-specific losses, because most Euryarchaeota, and all other members of the Asgard archaea, do encode glycerol-1-phosphate dehydrogenase (7).

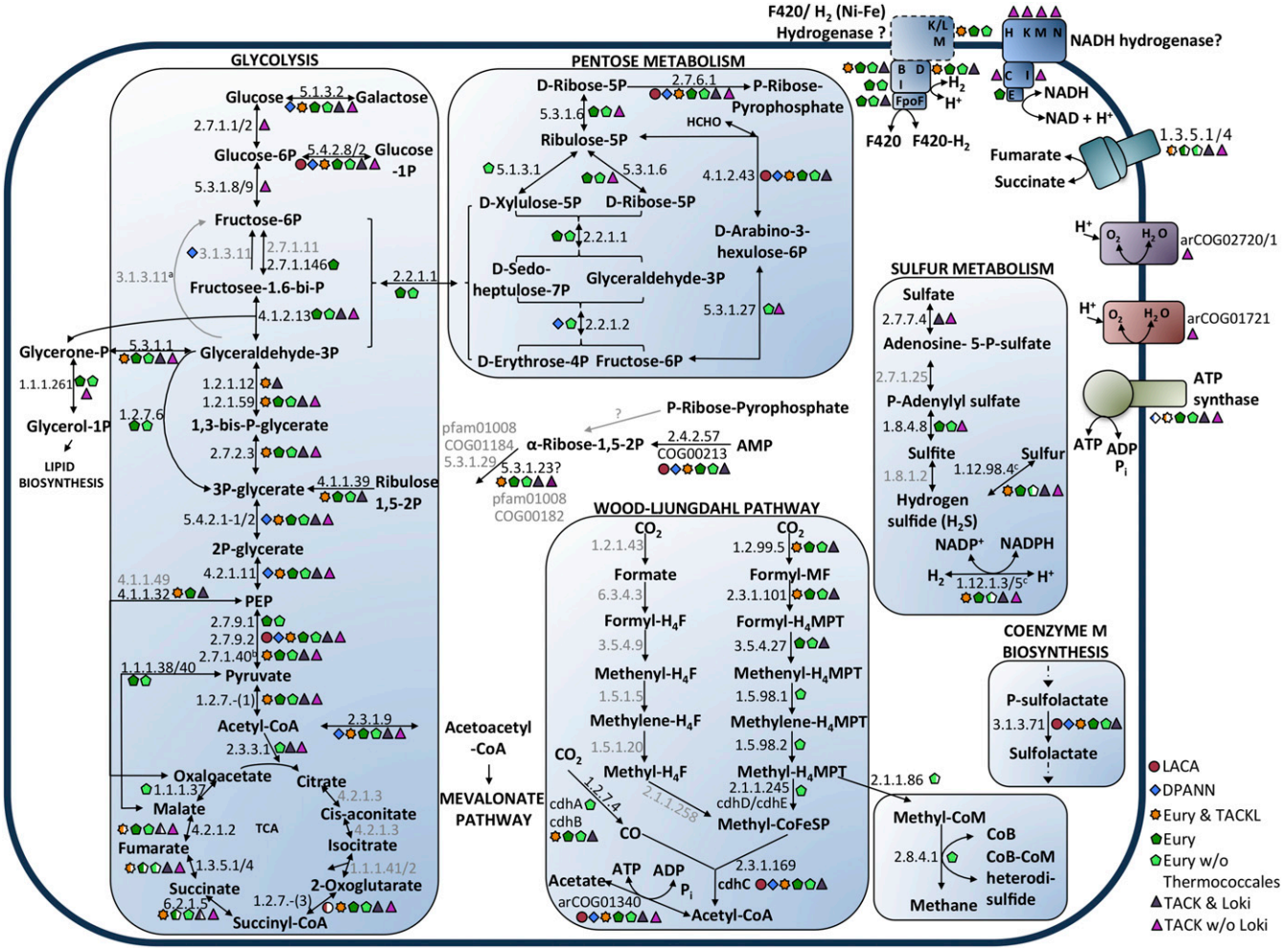

Fig. 4. Inference of ancestral archaeal metabolisms under the DTL model. The reconstruction is based on genes that could be mapped with $P > 0.5$ to a series of key nodes on the archaeal tree under the ML reconstruction of gene family evolution displayed in Fig. 3. The presence of a gene at a node is indicated by the symbols shown in the key, and partially filled symbols indicate that only some of the subunits composing a particular enzyme were present. Owing to the occasional extinction of gene families during evolution, as well as the increased uncertainty associated with DTL scenarios in the early regions of the tree, reconstructions of gene content at deeper nodes are increasingly incomplete. Nonetheless, the reconstruction supports the proposal that the ancestral archaeon was an anaerobe that encoded a subunit (cdhC) of CO dehydrogenase/acetyl-CoA synthase, the key enzyme of the Wood–Ljungdahl pathway. Aerobic metabolisms evolved later and independently in several different archaeal lineages, perhaps associated with the rise in atmospheric oxygen that began 2.5–2.3 Gya (82). Eury, Euryarchaeota including Thermococcales; Eury w/o Thermococcales, Euryarchaeota without Thermococcales; TACKL, TACK and *Lokiarchaeum*; B, nuoB/Ni Fe-hydrogenase III small subunit/coenzyme F420-reducing hydrogenase, gamma subunit; D, nuoD/Ni Fe-hydrogenase III large subunit and subunit G/coenzyme F420-reducing hydrogenase, alpha subunit; FpoFm coenzyme F420-reducing hydrogenase, beta subunit. [a]The bifunctional fructose-1 6-bisphosphate aldolase/phosphatase FBPA/FBPase (arCOG04180) (98) was not predicted to be present in any of the ancestors. [b]Pyruvate kinase is a glycolytic enzyme only. [c]A tetrameric protein complex with α, δ, β, and γ subunits, which in *Pyrococcus* functions as both a sulfur reductase (α, δ) and a hydrogenase (β, γ) (99); the ancestral enzyme also might have been bifunctional.

Moving beyond the LACA, our analyses suggest that the Euryarchaeota/TACK common ancestor was also an anaerobe, possessing enzymes including superoxide reductase/desulfoferredoxin (pfam01880) commonly found in modern anaerobic and microaerophilic organisms. This ancestor also might have possessed an anaerobic proton-pumping system comprising membrane-bound F420- and/or H_2-dependent hydrogenases. Some of the recently discovered anaerobic Archaea that branch near the base of Euryarchaeota or TACK, such as *Lokiarchaeum* (6, 77), the Hadesarchaea (78), and some Bathyarchaeota (79), also have retained genes of the Wood–Ljungdahl pathway. Whereas some of the key enzymes of methanogenesis could be mapped (with $P > 0.5$) only to the base of the Euryarchaeota (Fig. 4), the recent discovery of methyl-CoM reductase in large-genome Bathyarchaeota (79) is consistent with an early origin of methane metabolism in Archaea (79), as is evidence of the presence of microbial methane—today produced exclusively by Archaea—in 3.46-billion-y-old rocks (19, 80).

Our analyses indicate that oxidative phosphorylation as attested by terminal oxidases and NADH dehydrogenase appears to have been acquired independently in several descendent lineages, including the TACK Archaea after their divergence from *Lokiarchaeum* and the stem leading to the Haloarchaeota (81). It is tempting to speculate that these parallel acquisitions of oxidative metabolisms may have been associated with the rise in atmospheric oxygen beginning around 2.5–2.3 billion y ago (82). Some of the genes today involved in sulfur metabolism also appeared first in the Euryarchaeota/TACK ancestor, including a potential sulfhydrogenase. Others, particularly genes for sulfur reduction, appear to have originated independently along the stems leading to different crenarchaeotal and euryarchaeotal lineages.

The inferred metabolic map of the DPANN common ancestor is similar to that of the LACA, consistent with the anoxic environments from which many members of this lineage have been obtained (5). However, in contrast to the LACA, the DPANN ancestor also encodes additional components of central metabolism, including enzymes involved in glucose and pentose sugar metabolism. The reconstruction suggests that the DPANN common ancestor may have been capable of anaerobic proton pumping via a V-type ATP synthase, given that two subunits of this membrane complex were mapped to the root of DPANN (*SI Appendix*, Table S6). Some modern DPANN species have subsequently lost these subunits, and it has been suggested (5, 83) that these may have a fermentative, parasitic, or symbiotic lifestyle.

Ancestral Growth Temperatures. Previous work exploiting the correlation between sequence composition and optimal growth temperatures (OGTs) suggested that early Archaea were (hyper)thermophiles, with mesophily arising more recently in archaeal evolution (84, 85). Given that some DPANN genomes have been obtained from mesophilic environments, we investigated the impact of a basal DPANN clade on estimates of ancestral temperature. Our 45-gene alignment displayed a strong correlation between amino acid composition and OGT for modern Archaea (*SI Appendix*, Fig. S18), allowing us to infer temperature optima for ancestral nodes in the tree. We sampled 100 ancestral sequences for each node at the base of the tree using the branch-heterogeneous CoaLA model (84), performing the analysis both with and without DPANN (*Materials and Methods*). The LACA and the last common ancestors of each of the major archaeal clades (DPANN, Euryarchaeota+TACK/*Lokiarchaeum*, Euryarchaeota, and TACK+*Lokiarchaeum*) were all inferred to be thermophiles, and these inferences were robust to the inclusion of DPANN in the analysis (*SI Appendix*, Table S7); the median optimal growth temperature estimate for the LACA was 73.1 °C in the full analysis, and 75.7 °C in the analysis without DPANN. Interestingly, our model predicts mesophilic optimal growth temperatures for most modern DPANN genomes, consistent with the idea (84, 85) that adaptation to mesophily from a thermophilic ancestor occurred independently in each of the major archaeal clades.

Inferring Ancestral Genome Sizes. The DTL model provides inferences of ancestral genome size, and, because the reconciliation model explicitly allows for horizontal transfer as well as gene loss, there is no trend toward inferring increasing genome size for earlier nodes on the tree. Thus, the use of this model ameliorates the "genome of Eden" (86) problem, a tendency toward inferring unrealistically large ancestral genomes in the absence of HGT that is so marked that it has been used to set a lower bound on rates of HGT through time (87). Previous simulation studies (67) and analyses of empirical data (88) have suggested that ancestral gene content inferences under this model are realistic and robust to gene tree uncertainty, and thus the ancestral sizes that we present here have been corrected to account for gene families that have been lost in all sampled species, as described above. Our analyses suggest that there has been an ongoing increase in gene content throughout archaeal history, from ∼1,090 genes in the common ancestor to 537–5,359 (mean, 1,686.4) genes among modern lineages. This trend is not dependent on the basal placement of the DPANN clade in the tree; in the analysis without DPANN, the common ancestor was predicted to encode 1,328 genes, increasing to 1,373–5,359 (mean, 2,081.1) genes among modern genomes. These reconstructions do not support the hypothesis, based on an analysis of phylogenetic presence-absence profiles (89), that a large-genome archaeal common ancestor gave rise to modern lineages by genomic streamlining.

Dynamics of Archaeal Genome Evolution: Gene Transfers, Duplications, and Losses. Our reconciliations suggest that archaeal gene family evolution has been largely vertical (see also ref. 26), because for the majority (15,623) of gene families, vertical transmission events outnumber horizontal transfers [transfer ratio (TR) <0.5] (*Materials and Methods*). Interestingly, the distribution of TRs is multimodal, with a small peak of genes at TR >0.5 (*SI Appendix*, Fig. S19). In agreement with previous work on the transferability of genes with different kinds of functions (90), functional category had a significant effect on TR ($P = 7.26 \times 10^{-134}$, ANOVA), with genes involved in carbohydrate metabolism (COG category G; $P = 2.5 \times 10^{-10}$, Fisher's exact test) and defense functions (COG category V) enriched in the set of frequently transferred genes with TR >0.5 ($P = 6.7 \times 10^{-12}$, Fisher's exact test). Despite the overall predominance of vertical inheritance and the observed functional biases associated with HGTs, the cumulative effect of HGT on archaeal genomes is striking, and HGTs outnumber gene duplications on most (96 of 119) branches. Note that our inferences regarding HGT may represent underestimates, because increased taxon sampling may suggest that some inferred duplications instead are HGTs among close relatives. Remarkably, our reconstruction suggests that all of the gene families present at the root have experienced at least one HGT during archaeal evolution. Only 136 archaeal gene families are inferred to have entirely escaped HGT, and these are all recent originations.

In the discussion that follows, we define gene acquisitions as the sum of new genes that arise on a branch owing to lineage-specific innovation (i.e., apparently new genes with no detectable similarity to sequences from outside the subtended clade) or are obtained by HGT. The distributions of gene acquisition, duplication, and loss rates are continuous and correlated across the archaeal tree (all correlation coefficients ≥0.31; $P < 0.01$) (*SI Appendix*, Fig. S20), in agreement with some previous analyses of archaeal genome evolution (91). Acquisitions and duplications also are positively correlated with branch lengths ($P < 0.05$; *SI Appendix*, Fig. S20). Thus, according to our analyses, archaeal evolution is generally characterized by steady rather than punctate rates of genome change, with more events occurring on longer branches of the tree. Nonetheless, distributions for all of these processes have clear outliers (Fig. 5), indicating that some branches on the archaeal tree are exceptions to these background rates (*SI Appendix*, Table S8). The branch with the greatest number of gene acquisitions and duplications is that leading to the composite *Lokiarchaeum* genome. The high numbers may reflect the origin of some duplicated gene families shared with eukaryotes in the common ancestor of both lineages. However, the redundancy of the *Lokiarchaeum* composite genomic bin, which is estimated to include contigs from 1.4 closely related strains, also may be inflating estimates (6). The tip lineage with the second-highest number of acquisitions is that leading to *Nitrososphaera gargensis,* a group 1.1b Thaumarchaeote inhabiting a hot spring environment (92). This metabolically versatile archaeon has a much larger genome size (2.83 Mb) and gene complement (3,566 ORFs) than group 1.1a Thaumarchaeota, and has been inferred to have undergone extensive gene duplication, de novo gene origination, and horizontal acquisition (93). The crenarchaeote *Sulfolobus solfataricus* and the euryarchaeote *Methanomassiliicoccus luminyensis* also have experienced a large number of gene acquisitions, in both cases including a range of mineral and nutrient transporters.

The two stem lineages in which we observe the greatest number of gene acquisitions are the branches leading to the Haloarchaea and the Thaumarchaeota, two lineages that have undergone significant ecological transitions. Haloarchaea are suggested to have evolved into oxygen-respiring, light-harvesting heterotrophs from a methanogenic ancestor (81), whereas Thaumarchaeota may have evolved an ammonia-oxidizing lifestyle from an anaerobic ancestor (94). Horizontal transfer of metabolic genes from Bacteria has been implicated as an important process in these transitions (94–96), although the number of inferred transfers is sensitive to both

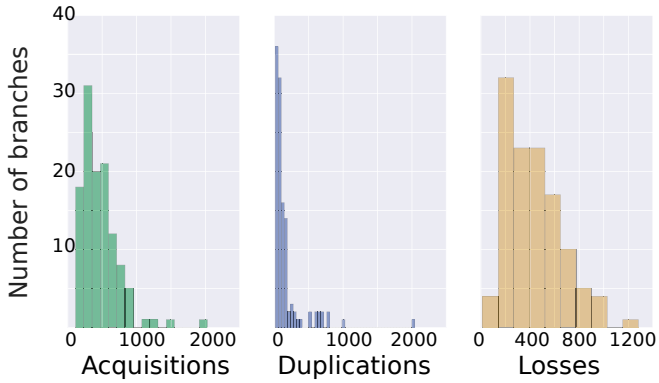

Fig. 5. Distributions of gene acquisition, duplication, and loss rates across the archaeal tree. We observed clear outliers for each distribution. The greatest number of outliers correspond to the branch leading to *Lokiarchaeum* (gene duplications) and to the branches leading to the Haloarchaea (gene acquisitions and gene losses) and Thaumarchaeota (gene acquisitions).

the method used for mapping (26, 91, 94) and the taxonomic sampling of the lineages involved (97). Because the taxon sampling in our study was optimized for rooting the entire archaeal tree, our sampling within each of these groups is limited (two Haloarchaea, four Thaumarchaeota). Thus, from our analyses, it is difficult to pinpoint when these transfers occurred during haloarchaeal and thaumarchaeotal evolution.

Although we infer different numbers to published analyses, our findings are consistent with a substantial number of functionally relevant HGTs among Bacteria, Haloarchaea, and Thaumarchaeota. These include components of the electron transport chain and membrane transporters in Haloarchaea (*SI Appendix*, Table S9) and key components of the ammonia-oxidizing machinery in Thaumarchaeota (*SI Appendix*, Table S10), in agreement with recent large-scale phylogenetic analyses of genes shared between bacteria and mesophilic archaea (94–96). Interestingly, both stems also are inferred to have experienced a relatively large number of de novo gene originations and expansions of ancestral archaeal families. In the case of Haloarchaea, we identified 379 expanded or acquired gene families, including 109 (29%) corresponding to de novo gene origins, 156 (41%) corresponding to expansions of ancestral archaeal gene families, and 114 (30%) potential interdomain HGTs. These families have homologs in bacteria, although resolving the direction of transfer is difficult given the present data and methods. For Thaumarchaeota, we identified 17 de novo origins (16%), 72 expansions (69%), and 15 interdomain HGTs (15%). The haloarchaeal stem was the branch experiencing the greatest number of gene losses.

Conclusion. In the present study, we used large amounts of genomic data and a method that implicitly considers patterns of genome duplication, HGT, and gene loss (67) to generate a rooted tree for the Archaea, one of the two primary domains of life (19). The DTL model performed well in simulations and in our case used phylogenetic signals from 31,236 homologous gene families, compared with the small universal core of single-copy orthologous genes typically used for outgroup rooting. The DTL analyses infer a new root between the DPANN clade and all other Archaea, with the Euryarchaeota and the TACK/*Lokiarchaeum* clade resolved as monophyletic sister lineages. Monophyly of DPANN was supported by supertrees, supermatrices, and DTL modeling, and thus, notwithstanding legitimate concerns about potential LBA artifacts, is the hypothesis best supported by our analyses. Its robustness will be tested as methods and genomic sampling of the relevant groups continue to improve.

The DTL analysis and new root provides inferences of gene content evolution that are consistent with inferences of early archaeal physiology based on other lines of evidence (20, 72, 73). Our analysis suggests that the LACA was an anaerobe that fixed carbon via the Wood–Ljungdahl pathway, and that adaptations to aerobic metabolism evolved independently across the tree. We infer that ecological transitions within the Archaea are associated with substantial gene content turnover, involving both HGT and the evolution of lineage-specific genes. Although our analyses agree that HGT is an important feature of archaeal evolution, the majority of transmission events appear to be vertical rather than horizontal, preserving a strong vertical trace between lineages. In contrast to hypotheses in which a large-genomed archaeal common ancestor gave rise to modern lineages by streamlining (89), the DTL analyses imply a moderate increase in gene content throughout archaeal history from a common ancestor that had a relatively small genome. Our analyses also suggest that adaptation to mesophily from a thermophilic ancestor occurred independently in each of the major archaeal clades.

Materials and Methods

Sequences and Alignments. We used the OMA algorithm (22) to identify orthologous gene families on 62 archaeal genomes, resulting in 4,664 orthologous families with at least four members. Families were screened for interdomain HGT using a BLASTP-based protocol requiring that all genes be more similar (lower E-value) to one or more sequences from other archaeal genomes than to any sequences from bacteria or eukaryotes. Sequences that did not meet this requirement were filtered out, resulting in a set of 3,266 orthologous archaeal protein families containing four or more sequences. Sequences were aligned using MUSCLE (100), and poorly aligning regions were identified and removed using BMGE (47) under the BLOSUM30 substitution matrix.

Single Gene Trees. Single gene trees were inferred using the C60+LG model in PhyloBayes 4.1 (101); this model is optimized for smaller datasets on which more general models (102) can show convergence problems. Two chains were run in parallel, and convergence was assessed using the bpcomp and tracecomp programs in PhyloBayes. A consensus tree was built once the maximum interchain discrepencies in bipartition frequencies and a range of continuous model parameters had dropped below 0.1, with effective sample sizes for continuous parameters >100. One-quarter of sampled points were discarded as burn-in.

Supermatrices. We identified 57 orthologous families that were present on all, or all but one, of the completely sequenced genomes in our dataset. This slightly relaxed criterion allowed for occasional gene losses or misannotations on published genomes. Because assemblies for uncultured organisms are often incomplete, we did not require the presence of these marker genes on all single-cell genomes, but did include orthologs where present. The gene trees from these families were visually inspected, and only those recovering the monophyly of both the TACK Archaea and the core Euryarchaeota at PP >0.7 were concatenated. This resulted in a final set of 45 single-copy orthologous marker genes and a concatenation of 10,738 aligned amino acid positions. The LG, CAT+Poisson, and CAT+GTR substitution models were fit to this concatenation using PhyloBayes-MPI (103), with posterior predictive simulations used to evaluate model adequacy. None of these models was able to adequately account for the across-site ($z = 6.14$; $P = 0$, CAT+GTR) or across-branch compositional heterogeneity ($z = 8.001$; $P = 0$, CAT+GTR), potentially leading to phylogenetic artifacts. Thus, we explored data-recoding techniques as a means of ameliorating these compositional biases. Even after data recoding, the data contained both across-site and across-branch compositional heterogeneity that was not adequately anticipated by the model ($P = 0$, CAT+GTR+Dayhoff4), but a reduction in the z-scores associated with the posterior predictive tests ($z = 3.43$ for across-site compositional heterogeneity, $z = 4.66$ for across-branch compositional heterogeneity) suggested improved model fit.

Supertrees. We used MRP (25) to infer a supertree from majority-rule posterior consensus trees for the orthologous archaeal protein families. Out of 3,266 families, 24 produced comb trees and 3,242 had at least some resolution at PP >0.5; we used this latter set in the MRP analysis. The input trees were definitive, producing a single most parsimonious supertree.

Modeling Gene DTL. We built an expanded set of gene family trees that included both paralogs and orthologs. We performed an all-versus-all BLASTP

analysis of our 62 archaeal genomes (E-value threshold $<10^{-5}$), then inferred gene clusters using MCL (Markov clustering algorithm) with an inflation parameter of 1.4. Sequence sets were aligned and masked, and gene trees were inferred as above. Because these homologous gene families were larger than the orthologous gene families used above, we used slightly less stringent convergence criteria to obtain gene tree samples within a tractable amount of time (maximum difference in bipartition frequencies, 0.3; minimum effective sample sizes of continuous parameters, >50). We performed gene tree-species tree reconciliation using the ALEml_undated algorithm of the ALE package (67) (https://github.com/ssolo/ALE), which uses a probabilistic approach to exhaustively explore all reconciled gene trees that can be amalgamated as a combination of clades observed in a sample of gene trees. We estimated a single global set of ML DTL rates for each rooted species tree. In estimates of ancestral gene content, we used extinction probabilities conditional on the estimated ML rates to account for genes that have gone extinct.

Functional Annotation of Gene Families. ArCOG categories were assigned by BLASTing each member of each gene family against the 2015 version of the ArCOG database (104). For KEGG Orthology number assignment, we selected the medoid (the sequence with the shortest summed genetic distances to all other sequences in the family, calculated under the BLOSUM62 substitution matrix) for annotation, and searched this against the KEGG database (105) using the KAAS annotation server (106) (accessed February 2016). We reconstructed ancestral gene family repertoires from the DTL model by selecting all families predicted to be present at a given node with $P \geq 0.5$. We assessed the metabolic capabilities of ancestral genomes using the KEGG Module tool. Genes gained along specific branches of the archaeal tree were identified by screening for gene families whose size increased by ≥ 1 along that branch. The origins of these genes were assessed by BLASTing against our sample of archaeal genomes and a reference set of bacterial genomes. Genes gained on a branch with significant hits to one or more bacterial genomes (BLASTP E-value <0.00001) but no homology to any other archaeal genomes were classified as putative bacterial HGTs, genes with no homology to any other sequenced genome were classified as de novo gene originations, and genes with homology to other archaeal genomes were classified as expansions of ancestral archaeal families, whether by gene duplication or by within-Archaea HGT transfer.

Ancestral Temperature Estimates. Estimates of ancestral OGTs are based on reconstructed ancestral sequences based on three different concatenates, analyzed under the LG+4G+COaLA model. At the root, 19 free parameters were estimated for the amino acid frequencies. Parameters were estimated by ML using the Bio++ libraries (107). A correspondence analysis on amino acid composition was performed for each of the three concatenates using the R package ADE4 (108). In all cases, axis 2 was found to correlate with OGT. The OGT for ancestral nodes was predicted by linear regression (predict.lm in R, with "interval= confidence") to take uncertainty in the parameter estimates of the regression line into account.

ACKNOWLEDGMENTS. T.A.W. is supported by a Royal Society University Research Fellowship. T.M.E. acknowledges support from the European Research Council Advanced Investigator Programme and the Wellcome Trust (Grants ERC- 2010-AdG-268701 and -045404). This work also was supported by grants from the European Research Council (ERC Starting Grant 310039-PUZZLE_CELL), the Swedish Foundation for Strategic Research (Grant SSF-FFL5), and the Swedish Research Council (Grant 2015-04959, to T.J.G.E.). A.S. received Marie Curie Intra-European Fellowship Grant 625521 from the European Union to join the T.J.G.E. laboratory. P.G.F. was supported by the Biotechnology and Biological Resources Sciences Research Council (Grant BB/G024707/1). B.B. was supported by the French Agence Nationale de la Recherche through Grant ANR-10-BINF-01–01, "Ancestrome". G.J.S. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under Grant Agreement 714774.

1. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87: 4576–4579.
2. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6:245–252.
3. Pester M, Schleper C, Wagner M (2011) The Thaumarchaeota: An emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol* 14:300–306.
4. Rinke C, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437.
5. Castelle CJ, et al. (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* 25:690–701.
6. Spang A, et al. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521:173–179.
7. Zaremba-Niedzwiedzka K, et al. (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541:353–358.
8. Guy L, Ettema TJG (2011) The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol* 19:580–587.
9. Petitjean C, Deschamps P, López-García P, Moreira D (2014) Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol Evol* 7:191–204.
10. Raymann K, Brochier-Armanet C, Gribaldo S (2015) The two-domain tree of life is linked to a new root for the Archaea. *Proc Natl Acad Sci USA* 112:6670–6675.
11. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA* 86:9355–9359.
12. Gogarten JP, et al. (1989) Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. *Proc Natl Acad Sci USA* 86:6661–6665.
13. Dagan T, Roettger M, Bryant D, Martin W (2010) Genome networks root the tree of life between prokaryotic domains. *Genome Biol Evol* 2:379–392.
14. Lake JA, Skophammer RG, Herbold CW, Servin JA (2009) Genome beginnings: Rooting the tree of life. *Philos Trans R Soc Lond B Biol Sci* 364:2177–2185.
15. Skophammer RG, Herbold CW, Rivera MC, Servin JA, Lake JA (2006) Evidence that the root of the tree of life is not within the Archaea. *Mol Biol Evol* 23:1648–1651.
16. Bergsten J (2005) A review of long-branch attraction. *Cladistics* 21:163–193.
17. Shavit L, Penny D, Hendy MD, Holland BR (2007) The problem of rooting rapid radiations. *Mol Biol Evol* 24:2400–2411.
18. Foster PG, Cox CJ, Embley TM (2009) The primary divisions of life: A phylogenomic approach employing composition-heterogeneous methods. *Philos Trans R Soc Lond B Biol Sci* 364:2197–2207.
19. Williams TA, Foster PG, Cox CJ, Embley TM (2013) An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504:231–236.
20. Sousa FL, Martin WF (2014) Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism. *Biochim Biophys Acta* 1837:964–981.
21. Sojo V, Pomiankowski A, Lane N (2014) A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol* 12:e1001926.
22. Roth ACJ, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 9:518.
23. Hrdy I, et al. (2004) Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* 432:618–622.
24. Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
25. Ragan MA (1992) Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol* 1:53–58.
26. Akanni WA, et al. (2015) Horizontal gene flow from Eubacteria to Archaebacteria and what it means for our understanding of eukaryogenesis. *Philos Trans R Soc Lond B Biol Sci* 370:20140337.
27. Pisani D, Cotton JA, McInerney JO (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol* 24:1752–1760.
28. Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM (2012) A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc R Soc B* 279:4870–4879.
29. Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM (2007) Of clades and clans: Tterms for phylogenetic relationships in unrooted trees. *Trends Ecol Evol* 22: 114–115.
30. Nunoura T, et al. (2011) Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* 39:3204–3223.
31. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA* 74:5088–5090.
32. Elkins JG, et al. (2008) A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci USA* 105:8102–8107.
33. Huber H, et al. (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 417:63–67.
34. Comolli LR, Banfield JF (2014) Inter-species interconnections in acid mine drainage microbial communities. *Front Microbiol* 5:367.
35. Youssef NH, et al. (2015) Insights into the metabolism, lifestyle and putative evolutionary history of the novel archaeal phylum Diapherotrites. *ISME J* 9:447–460.
36. Lasek-Nesselquist E, Gogarten JP (2013) The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol Phylogenet Evol* 69:17–38.
37. Williams TA, Embley TM (2014) Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol Evol* 6:474–481.
38. Saw JH, et al. (2015) Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. *Philos Trans R Soc Lond B Biol Sci* 370:20140328.
39. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27:401–410.
40. Philippe H, et al. (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol* 9:e1000602.
41. Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature* 440:623–630.

42. Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P (2005) Nanoarchaea: Representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol* 6:R42.
43. Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* 7:S4.
44. Ortiz-Alvarez R, Casamayor EO (2016) High occurrence of *Pacearchaeota* and *Woesearchaeota* (Archaea superphylum DPANN) in the surface waters of oligotrophic high-altitude lakes. *Environ Microbiol Rep* 8:210–217.
45. Schwartz RM, Dayhoff MO (1978) Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* 199:395–403.
46. Susko E, Roger AJ (2007) On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol* 24:2139–2150.
47. Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10:210.
48. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM (2008) The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci USA* 105:20356–20361.
49. Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol* 30:197–214.
50. Spang A, et al. (2013) Close encounters of the third domain: The emerging genomic view of archaeal diversity and evolution. *Archaea* 2013:202358.
51. Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* 47:9–17.
52. Hedtke SM, Townsend TM, Hillis DM (2006) Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol* 55:522–529.
53. Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol* 46:239–257.
54. Hirt RP, et al. (1999) Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci USA* 96:580–585.
55. Heaps SE, Nye TMW, Boys RJ, Williams TA, Embley TM (2014) Bayesian modelling of compositional heterogeneity in molecular phylogenetics. *Stat Appl Genet Mol Biol* 13:589–609.
56. Williams TA, et al. (2015) New substitution models for rooting phylogenetic trees. *Philos Trans R Soc Lond B Biol Sci* 370:20140336.
57. Philippe H, Laurent J (1998) How good are deep phylogenetic trees? *Curr Opin Genet Dev* 8:616–623.
58. He D, et al. (2014) An alternative root for the eukaryote tree of life. *Curr Biol* 24: 465–470.
59. Derelle R, Torruella G, Klime V (2015) Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci USA* 112:E693–E699.
60. Pisani D, et al. (2015) Genomic data do not support comb jellies as the sister group to all other animals. *Proc Natl Acad Sci USA* 112:15402–15407.
61. Morgan CC, et al. (2013) Heterogeneous models place the root of the placental mammal phylogeny. *Mol Biol Evol* 30:2145–2156.
62. Brown JR, Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci USA* 92:2441–2445.
63. Zhaxybayeva O, Lapierre P, Gogarten JP (2005) Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* 227:53–64.
64. Gogarten JP, Murphey RD, Olendzenski L (1999) Horizontal gene transfer: Pitfalls and promises. *Biol Bull* 196:359–361, discussion 361–362.
65. Abby SS, Tannier E, Gouy M, Daubin V (2012) Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci USA* 109:4962–4967.
66. Szöllosi GJ, Boussau B, Abby SS, Tannier E, Daubin V (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci USA* 109:17513–17518.
67. Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V (2013) Efficient exploration of the space of reconciled gene trees. *Syst Biol* 62:901–912.
68. Höhna S, Drummond AJ (2012) Guided tree topology proposals for Bayesian phylogenetic inference. *Syst Biol* 61:1–11.
69. Larget B (2013) The estimation of tree posterior probabilities using conditional clade probability distributions. *Syst Biol* 62:501–511.
70. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508.
71. Hug LA, et al. (2016) A new view of the tree and life's diversity. *Nat Microbiol* 1: 16048.
72. Sousa FL, Nelson-Sathi S, Martin WF (2016) One step beyond a ribosome: The ancient anaerobic core. *Biochim Biophys Acta* 1857:1027–1038.
73. Weiss MC, et al. (2016) The physiology and habitat of the last universal common ancestor. *Nat Microbiol* 1:16116.
74. Lombard J, López-García P, Moreira D (2012) The early evolution of lipid membranes and the three domains of life. *Nat Rev Microbiol* 10:507–515.
75. López-García P, Moreira D (2015) Open questions on the origin of eukaryotes. *Trends Ecol Evol* 30:697–708.
76. Villanueva L, Schouten S, Damsté JS (2017) Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the "lipid divide". *Environ Microbiol* 19: 54–69.
77. Sousa FL, Neukirchen S, Allen JF, Lane N, Martin WF (2016) Lokiarchaeon is hydrogen-dependent. *Nat Microbiol* 1:16034.
78. Baker BJ, et al. (2016) Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nat Microbiol* 1:16002.
79. Evans PN, et al. (2015) Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* 350:434–438.
80. Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y (2006) Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* 440: 516–519.
81. Nelson-Sathi S, et al. (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci USA* 109:20537–20542.
82. Bekker A, et al. (2004) Dating the rise of atmospheric oxygen. *Nature* 427:117–120.
83. Waters E, et al. (2003) The genome of *Nanoarchaeum equitans*: Insights into early archaeal evolution and derived parasitism. *Proc Natl Acad Sci USA* 100:12984–12988.
84. Groussin M, Boussau B, Gouy M (2013) A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst Biol* 62:523–538.
85. Groussin M, Boussau B, Charles S, Blanquart S, Gouy M (2013) The molecular signal for the adaptation to cold temperature during early life on Earth. *Biol Lett* 9: 20130608.
86. Doolittle WF, et al. (2003) How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Phil Trans R Soc B* 358:39–58.
87. Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104: 870–875.
88. Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B (2015) Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Phil Trans R Soc B* 370: 20140335.
89. Csurös M, Miklós I (2009) Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol* 26:2087–2095.
90. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–3806.
91. Groussin M, et al. (2016) Gene acquisitions from bacteria at the origins of major archaeal clades are vastly overestimated. *Mol Biol Evol* 33:305–310.
92. Hatzenpichler R, et al. (2008) A moderately thermophilic ammonia-oxidizing crenarchaeote from a hot spring. *Proc Natl Acad Sci USA* 105:2134–2139.
93. Spang A, et al. (2012) The genome of the ammonia-oxidizing *Candidatus Nitrososphaera gargensis*: Insights into metabolic versatility and environmental adaptations. *Environ Microbiol* 14:3122–3145.
94. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517:77–80.
95. Deschamps P, Zivanovic Y, Moreira D, Rodriguez-Valera F, López-García P (2014) Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota. *Genome Biol Evol* 6:1549–1563.
96. López-García P, Zivanovic Y, Deschamps P, Moreira D (2015) Bacterial gene import and mesophilic adaptation in archaea. *Nat Rev Microbiol* 13:447–456.
97. Becker EA, et al. (2014) Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS Genet* 10: e1004784.
98. Say RF, Fuchs G (2010) Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* 464:1077–1081.
99. Ma K, Schicho RN, Kelly RM, Adams MW (1993) Hydrogenase of the hyperthermophile *Pyrococcus furiosus* is an elemental sulfur reductase or sulfhydrogenase: Evidence for a sulfur-reducing hydrogenase ancestor. *Proc Natl Acad Sci USA* 90:5341–5344.
100. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
101. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.
102. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095–1109.
103. Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* 62: 611–615.
104. Makarova KS, Wolf YI, Koonin EV (2015) Archaeal clusters of orthologous Genes (arCOGs): An update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel)* 5:818–840.
105. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44:D457–D462.
106. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35: W182–5.
107. Guéguen L, et al. (2013) Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol* 30:1745–1750.
108. Dray S, Dufour AB (2007) The ade4 package: Implementing the duality diagram for ecologists. *J Stat Softw* 22:1–20.